Universität Bielefeld
Faculty of Technology – Semantic Computing Group

The USAGE review corpus for fine-grained, multi-lingual opinion analysis

Roman Klinger and Philipp Cimiano

Summary

- We make the "Bielefeld **U**niversity **S**entiment **A**nalysis Corpus for **G**erman and **E**nglish" publicly available
- Polarity, Subjectivity, Aspects and their Relations are annotated in ≈ 600 German and ≈ 600 English Amazon Reviews

Motivation

- Sentiment Analysis/Opinion Mining are important for a lot of domains
- Implementations typically trained and evaluated on manually annotated data
- Few German corpora on fine-grained sentiment analysis available
  - Classification, German Product reviews [1]
  - Layered, Subjectivity-focused [2]
- No German corpora for fine-grained sentiment analysis
- No German-English corpora for cross-lingual research

Research Questions

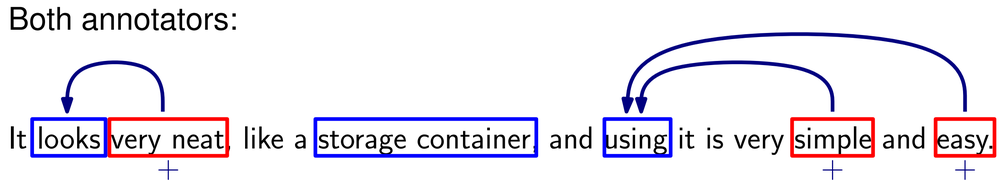
- How can we detect mentions of aspects and the corresponding evaluating phrases with their polarity?
- How can a model trained on the domain of a specific product be adapted to another domain with limited supervision?
- Can we exploit multilingual features to train sentiment analysis systems to improve performance?
- Can we train a model on one language and transfer that model automatically to another language?

Annotation

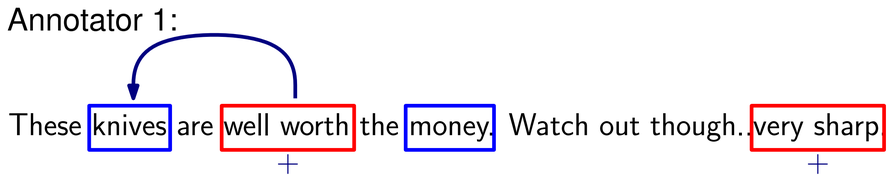
- Annotation scheme inspired by [3, 6]
- Reviews of Washing Machines, Microwaves, Vacuum Cleaners, Dish Washer, Toaster, Cutlery
- Annotation of Phrases and Tokens:
  - Subjective Phrases
    with Polarity being positive, negative or neutral
  - Aspects
    with additional information if corresponding to a different product
- Annotations of Relations:
  - Targets of Subjective Phrases
  - (some) Coreferences
- Every review annotated twice

Examples

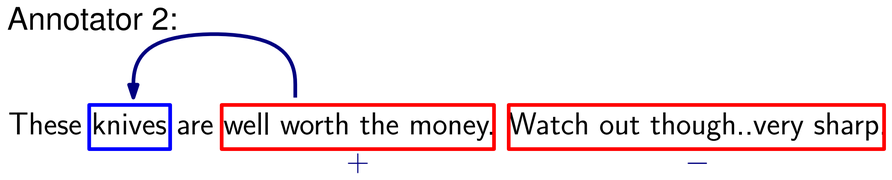
Both annotators:



Annotator 1:



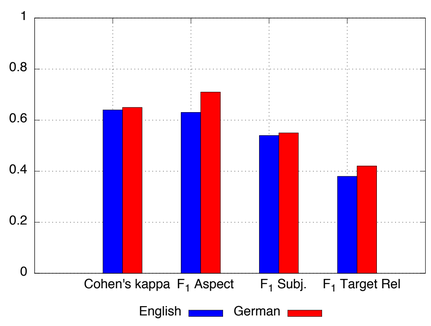
Annotator 2:



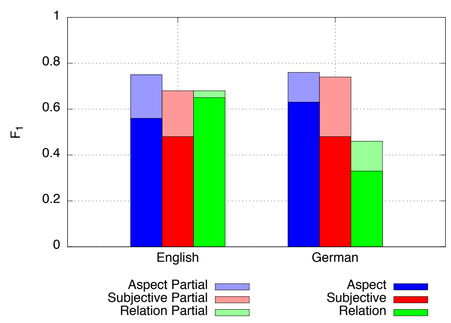
Statistics

	English	German
# reviews	622	611
# products	217	127
# Aspects	8545/6609	6340/5055
# Subj. +	3426/3600	3840/3717
# Subj. −	1799/1792	1094/1052
Target Rel.	4481/5180	4085/4643
Coref	67/462	37/224

Agreement



10-fold Cross-Validation



(Using the system described in [4, 5]), Annotator 1 only

Cross Domain

	English								German						
	coffee machine	cutlery	microwave	toaster	trash can	vacuum cleaner	washing machine	dish washer	Kaffeemaschine	Besteck	Mikrowelle	Toaster	Mülleimer	Staubsauger	Waschmaschine
Aspect	.50	.37	.50	.50	.45	.39	.50	.47	.53	.36	.48	.42	.43	.39	.43
≈	.69	.57	.70	.65	.62	.58	.63	.65	.63	.43	.59	.64	.55	.49	.57
Subjective	.50	.46	.49	.49	.50	.48	.45	.45	.46	.48	.52	.43	.44	.46	.42
≈	.70	.71	.68	.70	.69	.70	.66	.66	.74	.76	.73	.69	.69	.72	.69
Relation	.66	.68	.67	.62	.70	.60	.62	.64	.17	.20	.37	.15	.35	.19	.24
≈	.69	.71	.68	.65	.70	.64	.65	.66	.32	.30	.47	.15	.43	.37	.39

Availability

- Corpus available at dx.doi.org/10.4119/unibi/citec.2014.14
- `wget http://.../USAGE-corpus.tar.gz`
- `tar xzf USAGE-corpus.tar.gz`
- `cd USAGE-corpus/crawler`
- `mvn compile ; mvn assembly:single ; cd ..`
- `./crawler/bin/crawl.sh \`
  `files/en-coffeemachine.txt com`
  `files/en-coffeemachine-text.txt`

References

[1] K. Boland, A. Wira-Alam, and R. Messerschmidt. *Creating an Annotated Corpus for Sentiment Analysis of German Product Reviews*, volume 2013/05. GESIS Institute, 2013.

[2] S. Clematide, S. Gindl, M. Klenner, S. Petrakis, R. Remus, J. Ruppenhofer, U. Waltinger, and M. Wiegand. MLSA – A Multi-layered Reference Corpus for German Sentiment Analysis. In *Proceedings of the 8th International Conference on Language Ressources and Evaluation (LREC)*, pages 3551–3556, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[3] J. S. Kessler, M. Eckert, L. Clark, and N. Nicolov. The 2010 ICWSM JDPA Sentiment Corpus for the Automotive Domain. In *4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*, 2010.

[4] R. Klinger and P. Cimiano. Bi-directional Inter-dependencies of Subjective Expressions and Targets and their Value for a Joint Model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 848–854, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[5] R. Klinger and P. Cimiano. Joint and Pipeline Probabilistic Models for Fine-Grained Sentiment Analysis: Extracting Aspects, Subjective Phrases and their Relations. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, Dallas, TX, USA, 2013.

[6] D. Spina, E. Meij, M. de Rijke, A. Oghina, M. T. Bui, and M. Breuss. Identifying entity aspects in microblog posts. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 1089–1090, New York, NY, USA, 2012. ACM.

Contact: Roman Klinger, rklinger@cit-ec.uni-bielefeld.de

→ www.uni-bielefeld.de